# IEMS 402 - Statistical Learning

Practice Midterm, Spring 2024

**Name:** _____ **NetID:** _____

**While you wait, please read and check ☑ the following boxes:**

- ☐ Unless I have extra time with the academic accommodations, the time limit is **75 minutes**.
- ☐ I am taking this exam because I am a student enrolled in Professor Lu's section during this time. If this is not the case, I will leave the room immediately.
- ☐ I wrote my name and NetID (e.g. ab1234) at the top of this page.
- ☐ I will not detach any pages, especially not the scratch pages at the end.
- ☐ Except for multiple choice questions, I will show my work.
- ☐ If I need more space for an exercise, I will make a note and continue on one of the scratch pages.
- ☐ If I am caught in violation of academic integrity, including but not limited to peaking at another student's work, allowing another student to copy from my work, or speaking with another student, or using unauthorized resources, I will be asked to leave the exam and get a zero.


**Do not start the exam until you are permitted to.**

## Exercise I [15 points]

In many applications, labeled data is expensive and therefore limited, while unlabeled data is cheap and therefore abundant. For example, there are tons of images on the web, but getting labeled images is much harder. But what is the statistical value of having labeled data versus unlabeled data? This problem will explore this formally using asymptotics.

Specifically, suppose we have an exponential family model over a discrete latent variable $h$ and a discrete observed variable $x$:

$$p_\theta(h,x) = \exp\{\theta \cdot \phi(h,x) - A(\theta)\},$$

where $A(\theta) = \sum_{h,x} \exp\{\theta \cdot \phi(h,x)\}$ is the usual log-partition function.

Suppose that $n$ examples $(h^{(1)}, x^{(1)}), \ldots, (h^{(n)}, x^{(n)})$ are drawn i.i.d. from some true distribution $p_{\theta^*}$.

Define the following two estimators:

$$\hat{\theta}_{\text{sup}} = \arg\max_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(h^{(i)}, x^{(i)})$$

$$\hat{\theta}_{\text{unsup}} = \arg\max_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{h} p_\theta(h, x^{(i)}).$$

The supervised estimator $\hat{\theta}_{\text{sup}}$ uses the variable $h^{(i)}$ and maximizes the joint likelihood, while the unsupervised estimator $\hat{\theta}_{\text{unsup}}$ marginalizes out the latent variable $h$.

One important caveat: our results will hold when we assume that data is actually generated from our model family and that unsupervised learning is possible. Otherwise, labeled data is worth a lot more.

### a. (5 points) (supervised asymptotic variance)

Compute the asymptotic variance of $\hat{\theta}_{\text{sup}}$: that is, given that

$$\sqrt{n}(\hat{\theta}_{\text{sup}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_{\text{sup}}),$$

write an expression for $V_{\text{sup}}$ that depends on expectations/variances involving $\phi$.

**b. (5 points) (unsupervised asymptotic variance)**

Compute the asymptotic variance of $\hat{\theta}_{\text{unsup}}$: that is, given that

$$\sqrt{n}(\hat{\theta}_{\text{unsup}} - \theta^*) \xrightarrow{d} \mathcal{N}(0, V_{\text{unsup}}),$$

write an expression for $V_{\text{unsup}}$ that depends on expectations/variances involving $\phi$.

**c. (5 points) (comparing estimators)**

First, prove that $\hat{\theta}_{\text{sup}}$ has lower (or equal) asymptotic variance compared to $\hat{\theta}_{\text{unsup}}$.

## Exercise II [10 points]

The *adjacency matrix* $A$ of a graph $G$ on $n$ vertices is a symmetric $n \times n$ matrix whose entries are defined as $A_{ij} = 1$ if the vertices $i$ and $j$ are connected by an edge and $A_{ij} = 0$ otherwise.

Let us label the vertices of $G$ by the integers $1, \ldots, n$. A partition of the vertices into two sets can be described using a vector of labels

$$x = (x_i) \in \{-1, 1\}^n,$$

the sign of $x_i$ indicating which subset the vertex $i$ belongs to. For example, the three black vertices in Figure 3.10 may have labels $x_i = 1$, and the four white vertices labels $x_i = -1$. The cut of $G$ corresponding to the partition given by $x$ is simply the number of edges between the vertices with labels of opposite signs, i.e.

$$\text{CUT}(G, x) = \frac{1}{2} \sum_{\substack{i,j \\ x_i x_j = -1}} A_{ij} = \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - x_i x_j).$$

(The factor $\frac{1}{2}$ prevents double counting of edges $(i, j)$ and $(j, i)$.) The maximum cut is then obtained by maximizing $\text{CUT}(G, x)$ over all $x$, that is

$$\text{MAX-CUT}(G) = \frac{1}{4} \max \left\{ \sum_{i,j=1}^{n} A_{ij}(1 - x_i x_j) : x_i = \pm 1 \text{ for all } i \right\}.$$

Let us start with a simple 0.5-approximation algorithm for maximum cut — one which finds a cut with at least half of the edges of $G$.

**(a) (5 Points)** Partition the vertices of $G$ into two sets at random, uniformly over all $2^n$ partitions. Then the expectation of the resulting cut equals $0.5|E| \geq 0.5\,\text{MAX-CUT}(G)$, where $|E|$ denotes the total number of edges of $G$.

**(b) (5 Points)** For any $\epsilon > 0$ , give an $(0.5 - \epsilon)$-approximation algorithm for maximum cut, which is always guaranteed to give a suitable cut, but may have a random running time.

## Exercise III [20 points]

Let $X_1, \ldots, X_n \sim P$ where $X_i \in [0,1]$ and $P$ has density $p$. Let $\hat{p}$ be the histogram estimator using $m$ bins. Let $h = 1/m$. Recall that the $L_2$ error is

$$\int (\hat{p}(x) - p(x))^2 \, dx = \int \hat{p}^2(x) \, dx - 2 \int \hat{p}(x) p(x) \, dx + \int p^2(x) \, dx.$$

As usual, we may ignore the last term so we define the loss to be

$$L(h) = \int \hat{p}^2(x) \, dx - 2 \int \hat{p}(x) p(x) \, dx.$$

**(a) (5 Points)** Suppose we used the direct estimator of the loss, namely, we replace the integral with the average to get

$$\hat{L}(h) = \int \hat{p}^2(x) \, dx - \frac{2}{n} \sum_i \hat{p}(X_i).$$

Show that this fails in the sense that it is minimized by taking $h = 0$.

**(b) (5 Points)** Recall that the leave-one-out estimator of the risk is

$$\hat{L}(h) = \int \hat{p}^2(x)\,dx - \frac{2}{n}\sum_i \hat{p}_{(-i)}(X_i),$$

Show that

$$\hat{L}(h) = \frac{2}{(n-1)h} - \frac{n+1}{n^2(n-1)h}\sum_j Z_j^2,$$

where $Z_j$ is the number of observations in bin $j$.

**(c) (5 Points)** Show that $\hat{L}(h) - L(h) \xrightarrow{P} 0$.

## Exercise IV [5 points]

Explain the concepts: Curse of Dimensionality, Spurious local minima in Non-convex Optimization, Double Descent, Benigning Overfitting and Tempered Overfitting, Flat minimum

## Exercise V [20 Points]

For function class

$$\mathscr{F}_q = \{f_\theta(\cdot) = \langle \theta, \cdot \rangle : \theta \in B_q^d(R)\} \quad \text{with} B_q^d(R) := \left\{ \theta \in \mathbb{R}^d : \sum_{j=1}^d |\theta_j|^q \le R \right\},$$

(Here the dimension $d$ is allowed to be larger than the sample size $n$.) Consider $q = 1$ (i.e., $\ell_1$-constrained linear regression, or Lasso) and assume that the columns of $X$ are normalized to have $\ell_2$ norm bounded by $\sqrt{n}$. We'll try to show that

$$\log N(s, B_1^d(R), \|\cdot\|_2) \lesssim R^2 \left(\frac{1}{s}\right)^2 \log d.$$

We start the process from $B_1^{d,+} = \{x \in \mathbb{R}^d \mid \|x\|_1 \le 1 \text{ and } x_i \ge 0 \, \forall i\}$.

**(a)** For every $x \in B_1^{d,+}$ Find a probability distribution $\mathbb{P}(x)$ over $\{e_1, \ldots, e_d, 0\}$ such that

$$\mathbb{E}_{z \sim \mathbb{P}(x)}[z] = x$$

*hint*: $z = \sum_{i=1}^d x_i e_i + (1 - \|x\|_1) \cdot 0$

**(b)** We will draw $t$ samples $z_1, \ldots, z_t$ from the distribution where each $z$ is some $e_i$. After drawing the samples, we can take the average of the samples:

$$\bar{z} = \frac{1}{t} \sum_{i=1}^t z_i.$$

How large is $\mathbb{E}[\|\bar{z} - x\|_2^2]$?

**(c)** What's the covering number we can achieve if we create our $\epsilon$-cover using all the possible value of $\bar{z}$.

**(d)** The method we used here is called "empirical method of Maurey". Compared to volumetric estimate, show that empirical method of Maurey is better for large $\epsilon$ but worse for small $\epsilon$.

*(This question will not included in the midterm. This question needs first understand Question 2 from Homework 7)*

**Blank scratch page.**